Institute of Scientific Information of the Academy of Sciences of the USSR

V. P. CHERENIN and B. M. RAKOV

EXPERIMENTAL INFORMATION MACHINE of the INSTITUTE OF SCIENTIFIC INFORMATION of the USSR ACADEMY OF SCIENCES

MOSCOW, 1955

Institute of Scientific Information of the Academy of Sciences of the USSR

V. P. CHERENIN and B. M. RAKOV

EXPERIMENTAL INFORMATION MACHINE of the INSTITUTE OF SCIENTIFIC INFORMATION of the USSR ACADEMY OF SCIENCES

*

MOSCOW, 1955

INTRODUCTION

During the past two decades very many articles have appeared in the foreign press on the problem of the rapid searching of information. The first group of articles of this kind, quite considerable in number, were mainly figurative and popular expositions of the difficulties which led to this problem *. The doubling of the contents of libraries every 15 to 20 years, the ever increasing expenditures on the search for previously obtained data prior to any scientific research and many other facts aroused the apprehension that unless the methods of literature searching were radically improved, mankind would soon be flooded by the ever growing inflow of information. The suggestions put forth to meet this problem included mechanization of literature searching and the development of a suitable system of handling and recording the substance of the information and the question data.

The second group of articles published on this subject contained descriptions of such improved searching methods for individual limited branches of information. These included mainly punched cards with notched edges sorted manually by means of needles; superimposed punched cards, examined against the light; and punched cards, sorted on ordinary sorting machines **. The index data used for literature searching were bibliographical characterictics (year, country, author, etc.) and subject headings taken from preliminary compiled lists. For each particular case the articles would give these lists (at least partially), and would describe the method of encoding the characteristics and distributing them among the fields of the card (the card model).

These methods were found satisfactory for some individual limited branches of information and for individual use, but they could not be applied when it came to broader branches, due to

* Among Soviet authors, see Л. И. Гутенмахер. «К вопросу о машинной технике научной информации». Вестник Академии наук, № 8, 1952 год.

** For further information of the use of punch cards and bibliography, see R. S. Casey and J. W. Perry, «Punched cards», New York, 1951.

insufficient mechanization and, mainly, to insufficient standardization of characteristics and the difficulty of revealing the diversified relations between them. Thus, the general problem of literature searching fell into the problem of the *mechanization of searching* — the engineering of information machines and the problem of the *development of a special information (machine) language*, which would reflect all the relations between information characteristics, and the translation of information and question data into that language, i. e. *indexing*.

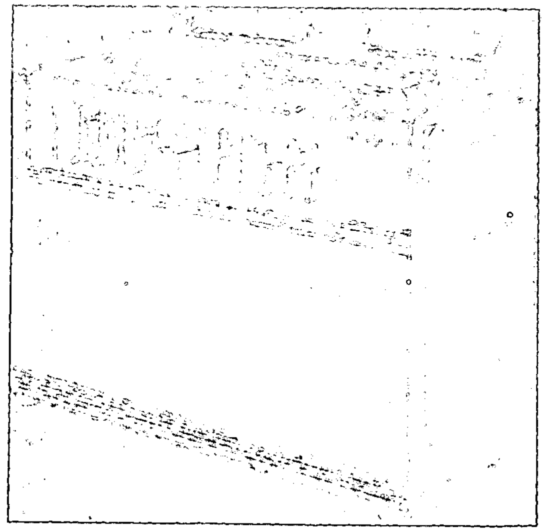

Fig. 1

If the development of an information machine for ordinary language is not practicable in the near future, this should not be very difficult for any not too complicated information language, especially in veiw of the achievements attained in the design of modern computing machines. It is natural, therefore, that nearly all the rest of the articles, not included in the first two groups, were devoted to the solution of this problem. There appeared descriptions of and brief information on the use of improved punched card sorting machines (the IBM Co. electronic statistical machine (fig. 1), J. Samain's first selector (fig. 2), etc.) and of specialized selecting machines (J. Samain's second Filmorex selector (fig. 3), the V. Buch microfilm selector (fig. 4), H. P. Luhn's machine). There were also indications as to the possibility of employing modern computers (UNIVAC, etc.) for literature searching purposes. Projects have been suggested of still more complicated and perfect information machines, capable of handling millions of information elements in a very short time. Just recently there appeared a communication * concerning a new very flexible system of information storage and searching developed by the Kodak Co., in which the information is recorded on 16×32 *mm* minicards.

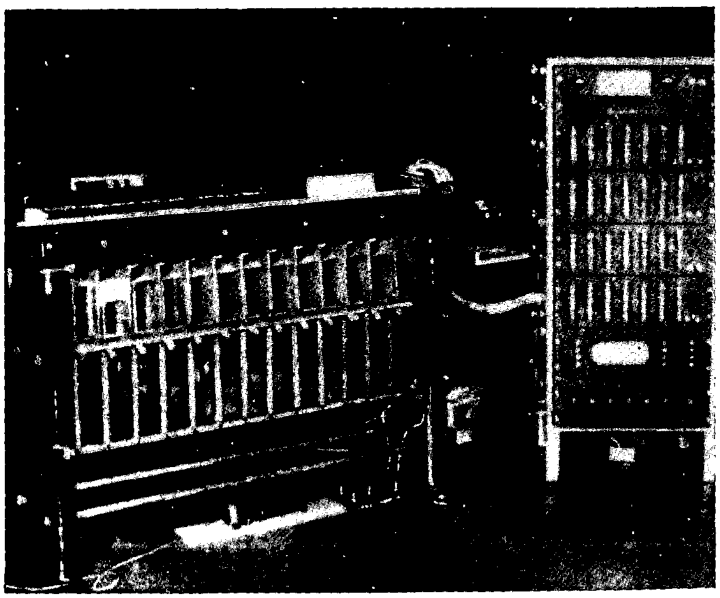

Fig. 2

It was much worse with the solution of the complicated problems, the development of a special information language and indexing. Only a small number of articles by M. Taube, J. W. Perry *et al.* were devoted to this problem. At the same time, the lag in the solution of these problems obstructed the effective mechanization of literature searching in the broad branches, even though highly efficient information machines already existed. All this somewhat sobered the over-optimistic adherents of such searches, who saw the solution of the entire problem almost exclusively in the development of a sufficiently efficient information machine,

* A. W. Tyler, W. L. Myers, J. W. Kuipers. «The Application of the Kodak Minicard System to Problems of Documentation». American Documentation, vol. VI, No. 1, 1955.

and made them look more attentively into the formulation of the problem *.

A careful examination of the problem leads to the following conclusions:

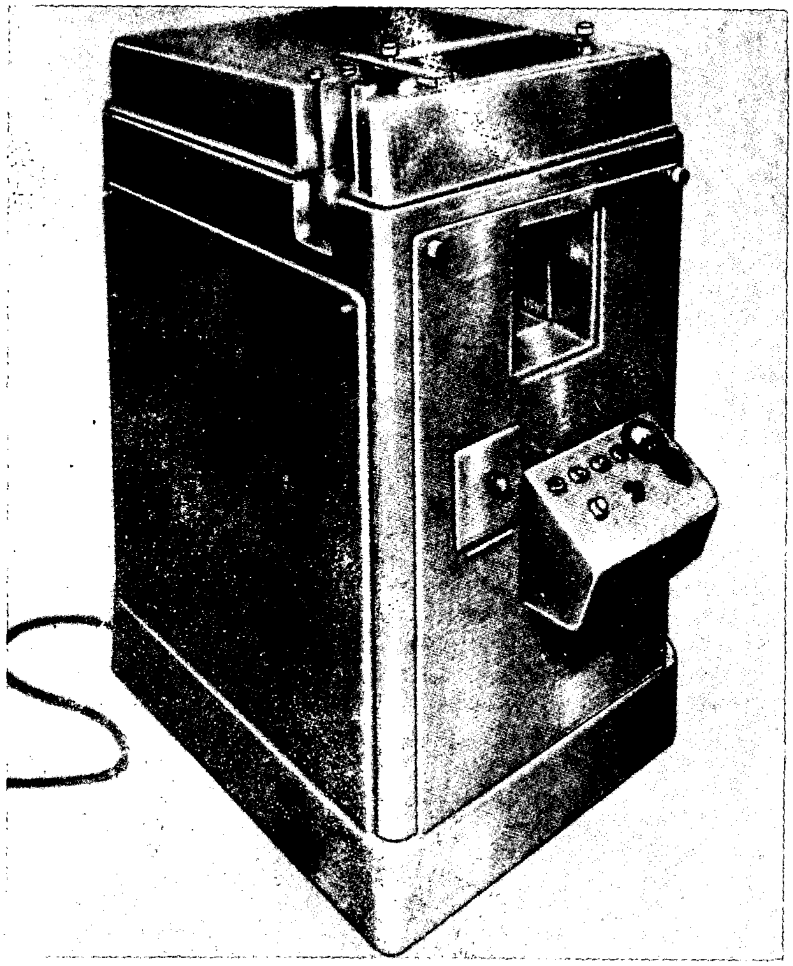

Fig. 3

* See series of articles by J. W. Perry et al. «Machine Literature Searching» in American Documentation for 1954 and 1955, and also: В. П. Черенин. «Некоторые проблемы документации и механизация информационных поисков», Москва, 1955 г.

1. The development of a special information language is the key problem among the three problems of literature searching in wide branches of information. Especially in changing over to machine searching the erudition of the searcher must be replaced by a reflection of the connection between the characteristics (the

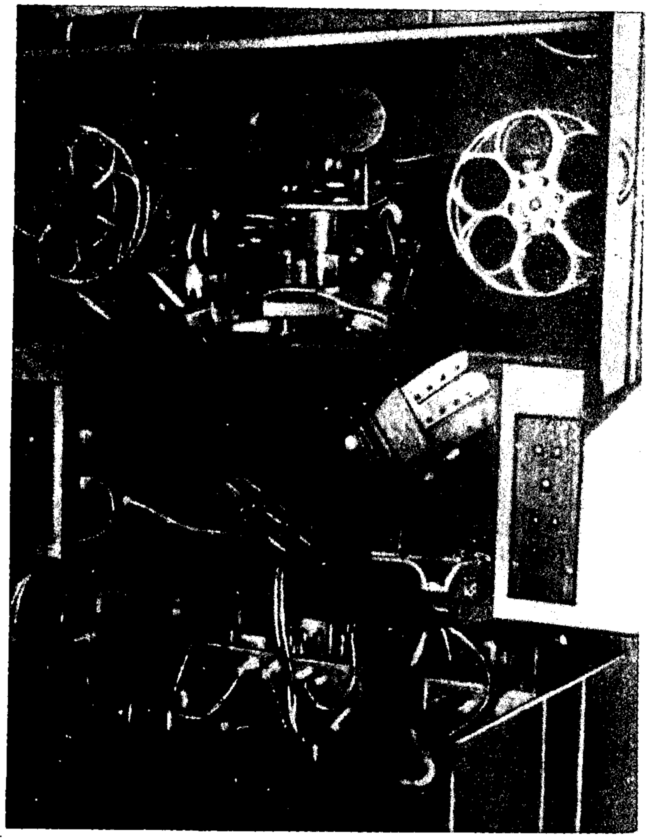

Fig. 4

words of the language) within the language Standardizing the characteristics and finding the relations between them is a very complicated job, closely connected with the work on terminology and classification, which can be carried out only by a large body of specialists. A uniform international information language would obviously be the most valuable in this connection

2. Certain particular branches of information, which have a more or less ready-made information language, can do quite well with relatively simple and inexpensive devices, such as notched punch cards, superimposed punch cards, improved sorters, Filmorex selectors, etc. More complicated and expensive machines should be built only for special information problems (with a known language), requiring high operativeness and searching speeds.

3. The machine to be used in developing an information language for wide branches of information and in the experimental work done in this connection must be sufficiently flexible and capable of passing easily and quickly from one language or code to another. Obviously, this machine need not be highly perfect, particularly as regards capacity and finished design.

The experimental information machine described below (EIM) was designed in the Institute of Scientific Information of the USSR Academy of Sciences precisely for experimental work in the development of an information language. Built on the basis of a conventional sorter with the least possible expenditures on reconstruction the machine, as will be demonstrated below, is capable of working with almost any language and code and is sufficiently flexible to carry out all kinds of experimental searches. Besides, after the test work has resulted in the development of a language and code for separate particular problems, the EIM can be modified without difficulty specially to solve these tasks and used for regular work in various establishments and enterprises, especially those already equipped with punch card machines.

---

1. Information Elements, Index Data, Conditions of Selection, Consecutive Action Machines *

Aggregates of information sought will be termed ***information elements***. The data in an information element are more or less closely related. Thus, for instance, one may assume that an information element is the aggregate of information contained in a certain article or abstract, or concerning a certain experiment, process, substance, plant, animal or person. Thus, the information element concept is a very conventional one, varying from case to case.

The index data or characteristic information of an information element to be used in searching fall into «*bibliographical*» and «*subject*» data. The former include all the data on the time, place, author and other conditions connected with the origin of the information element. The latter describe the subject matter. As a rule, each item of characteristic information is an inde pendent unit (conception) and may serve as question data, i. e. any one of them may become the basis for a search. In the simplest case the data contained in an information element can be described by merely naming the information characteristics without indicating any logical or grammatical relations between them within the information element. While this is quite natural for bibliographical characteristics, in the case of subject data, such a method often results in the construction of a rather complicated system of subject headings, which makes it difficult to standardize these headings and reveal the classification relations between them. Hence, part of the subject characteristics within the information element must be provided with the simplest types of grammatical relations: grouping of characteristics, similarly to the division of an ordinary article into sections, paragraphs and sentences; introduction of subordination of characteristics within groups, in analogy with the usual relations between the parts of

* For further information on the questions dealt with in the first four paragraphs, see the article by V. P. Cherenin cited on p.6.

a sentence, etc. The development of such a grammar is part of the work on the engineering of a special information language, and cannot be separated from the standardization of index data (the words of the language) and the ascertainment of the classification relations between them. Inclusion of grammatical relations does not require any considerable changes in the design of the machine dealt with in this article, and therefore, for the sake of simplicity we shall assume that the information elements can be defined by merely naming the aggregates of characteristics.

Question data are also represented either simply as aggregates of characteristics, or in the form of characteristics in some sort of grammatical relation with one another. For simplicity we shall assume that in this case also there are no grammatical specifications between the question characteristics. Then the basic condition for the selection of the necessary information element is that the question characteristics be included in the information element. However, a full inclusion of these characteristics may not always be required, and information elements may be sought which at least approximately correspond to the specified question. In this case the selection specification may be the presence in the information element of any m characterictics of the total number n contained in the question. More complicated selection conditions are also possible; for instance, one may require the selection of the information elements which contain among their characteristics A and B, or C, positively contain D, but should not contain E.

Thus, no matter what the selection specifications, the machine must compare the characteristics of the information element with those of the question specified. If the characteristics of the question are matched with the characteristics of many information elements at the same time the machine is a parallel action type; if the characteristics are matched consecutively with separate information elements, one by one, the machine is a consecutive action type. The consecutive action machine is capable of fuller comparisons, not only of the characteristics themselves, but of the grammatical relationships between them as well. Although the search rate of these machines is lower than that of the parallel action types, they require a minimum of equipment to effect the matching, since only one information element is compared with the question each time. The EIM machine is of the consecutive action type, and hence, parallel-action machines will not be discussed in this article.

Any consecutive action machine must contain:

a) a carrying or switching device to match the information elements one by one with the question characteristics;

b) a device for scanning the characteristics of the information element;

c) a storage unit to hold the characteristics of the question;

d) a device comparing question characteristics with those of the information element;

e) a selecting device, consisting of elements registering the results of matching and the fulfilment of the selection conditions, and of an executive organ which separates the information elements required from those not required.

In order that the machine should perform these operations it is necessary to render the characteristics both of the information element and the question in coded form to make them available for machine scanning. The coded index data of the information element are introduced into the machine on a so-called medium (film, magnetic tape, paper tape, punch cards, etc.), while the coded data of the question are kept in the storage device. Next to the symbol for the index data of the information element may be not only the name (number) of that information element, but also a full description of the data contained in the element in legible form (written by hand, typed or printed, in the form of a microfilm image, etc.). The sections of the medium corresponding to the required information elements are either separated out directly by means of the executive organ, as is the case with the punch cards in the EIM machine, or are photographed, or their content is re-recorded.

2. Coding

The codes discussed below may be applied in devices of any kind; however, having in mind the EIM, we shall deal specially with the possibility of representing index data as groups of holes punched in an ordinary 80 column card (fig. 5).

Each item of index data is represented singularly on the punched card and in the storage device, firstly by the configuration of the code symbols within the so-called field, and, secondly, by the location of this field. The assumed code symbols on the punch card are the holes, while in the storage unit they are the closed contacts of a plugboard or keyboard. We shall term these symbols secondary code symbols, forming a secondary coded designation of an item of index data in each field, in contradistinction to the primary code symbols and designations found in the vocabulary of the information language.

The primary code designations are indices and are, strictly speaking, the words of the information language. They represent regular sequences of primary code symbols, i. e. letters, digits, syllables, etc. written in one line. We shall term the spot occupied by the «*r-th*» primary symbol in the index, counting from left to right, the «*r-th*» rank. Let us assume that the total number of primary symbols which can occupy the «*r-th*» rank, increased by unity, if the absence of a symbol is not to be designated, is q . In particular, the indices may be words of an ordinary language in a definite grammatical form (for instance, in the Nominative case, Singular, etc.). Thus, the indices are intermediate forms between the usual denotation of characteristics and their secondary coded designations; having the clearness and legibility of the former, they at the same time quite reflect the latter, i. e. are in a way instructions used by the operator in entering the secondary code designations on the punch card and in the storage unit. Thus, for instance, if the punch card field consisting of one bit at the intersection of column 6 and line 9 is assigned to the item «automobile» according to a direct code (see below), the index may be «D, 6, 9». If the item is encoded according to

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

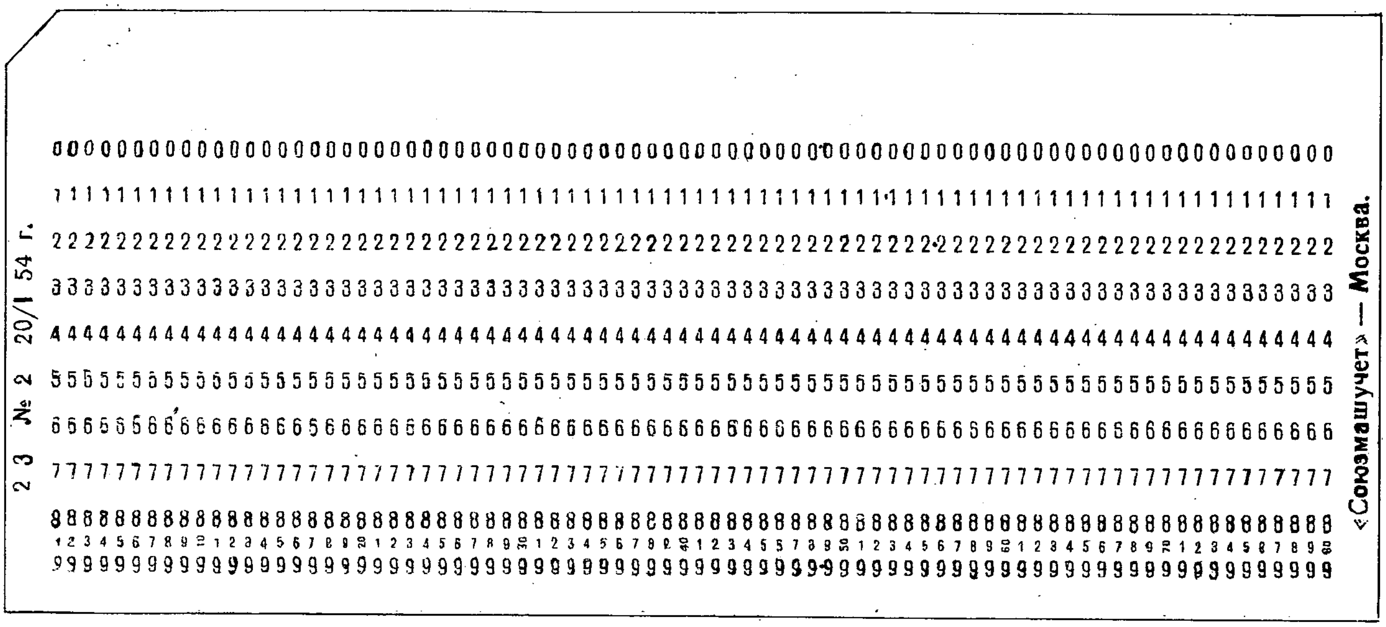

Fig. 5

a non-local alphabetic code (see also below), providing for entry of a 7-letter expression in each field, the index might be, for instance, «automob», etc.

Returning to the secondary code designations, let us see, first of all, how codes may be subdivided according to field location. Evidently, the following 4 cases are possible: the coded items differ only in the *configuration of the code symbols,* entered in a single field, common to all characteristics (superpositional code) or in separate fields for each characteristic, independent of the location of the latter (non-local code); the coded items differ only in the *location of the fields assigned to them* (direct code), only two configurations of symbols being possible in each field, to denote the presence or absence of the characteristic; and finally, the coded items may differ *both in field location and symbol configuration* (local code).

In the case of the *superpositional code,* when all the fields of the separate characteristics coincide, i. e. are superimposed, parasitic symbol configurations are apt to form, corresponding to characteristics which actually are not present in the information element. This leads to the selection of superfluous, unnecessary information elements when making searches, and hence, the use of this code is undesirable. In this case the characteristic question data are entered in the storage unit also by the superpositional code in a single common field. When matching (to see whether the question characteristics are or are not present among the index data of information elements) all those cards are selected which contain the specified plurality of holes among their holes.

When using a *direct* code each bit of the punch card (and each respective bit of the storage device) is made to correspond to a certain characteristic. The bit will be punched only on those cards which correspond to information elements containing the characteristic it represents. Since the total number of bits on the card is 960, the direct code may be used to record not more than 960 characteristics. It is advantageous when the information elements contain many characteristics out of a comparatively small aggregate total; otherwise the necessity of reserving place for the absent characteristics results in the punch card not being fully utilised. Obviously, in this case also, matching results in the selection of all the cards which contain the specified plurality of holes among their holes.

The direct code is a particular case of the *local* code. The latter is used when any characteristic may have not two values (present or absent), as is the case with the direct code, but many mutually exclusive values. For instance, the year or place of publication of an article are just such groups of mutually exclusive characteristics. Mutual exclusiveness of the characteristics means that any information element may contain not more than one characteristic of each group. A separate field of the punch card is assigned to each group of mutually exclusive characteristics, the location of this field determining the group, and the symbol configuration — the characteristic itself or its absence. The local code is advantageous when the information elements as a rule contain characteristics from almost each group. Then almost no fields are left unoccupied on the punch card. The question data are also recorded according to the local code. When matching, those cards are now selected on which the plurality of holes in the fields corresponding to those containing the question characteristic in the storage unit coincide with the plurality specified.

The *non-local* code is, for the majority of problems, not only the most highly manipulative, but also the most compact, since, unlike the local and direct codes, in using this code only those characteristics are entered and take up space on the card, which have been assigned to the given information element. The fields assigned to the individual characteristic items may be arranged arbitrarily on the punch card, as long as they can be scanned consecutively in matching. The question data are also entered according to the non-local code in separate fields of the storage unit. Each pair of characteristics is matched, one from the field of the punch card and one from the field of the storage unit, to reveal coinciding symbol configurations. Depending on the selection specifications, when matching, information elements may be selected containing all the characteristics specified in the question data, or a certain predetermined number of them. The number of pairs to be matched obviously equals the product of the number of characteristics on the punch card by the number of characteristics in the storage unit.

All the codes described may also be used simultaneously; in this case the card and the storage unit are subdivided into superpositional, direct, local and non-local code zones. Mixed codes, consisting of the main codes mentioned above, are also possible. For instance, to combine characteristics into groups (page 9) a superpositional non-local code may be used, where several characteristic items are entered according to the superpositional code in each of the fields mapped out for characteristic groups according to a non-local code. A superpositional local code is also possible when the characteristics are not entirely exclusive, and a local-non-local code, when the characteristics, entered according to the non-local code are preliminarily grouped, each group having its own zone of the non-local code, etc. The EIM can operate with all these codes, with one limitation concerning the configuration of the symbols, which will be described below.

In encoding the characteristics within the fields (in the case of the superpositional, local and non-local codes) by means of various symbol configurations, the following two possibilities should be distinguished above all:

a) all the configurations for a certain field consist of an equal number of symbols (*selector* code) and

b) different configurations within any field may consist of different numbers of symbols (*ordinal* codes, particularly, binary codes).

Application of the selector code, as we shall see below, considerably simplifies the matching unit, and for this reason the EIM uses only the selector code. This is the limitation mentioned above.

Since the general relationships between characteristics may be reflected in recording indices, and taking into account the necessity of facilitating the operations of encoding and decoding, it is advisable to encode the indices rank by rank. For this purpose the field of the punch card and of the storage unit is divided up into sections, the number of which equals the maximum number of ranks the indices entered in that field may have. A mutually single valued correspondence is established between the sections and the ranks. Then each section is subdivided into bits, in each of which one secondary code symbol can be entered. The number of bits in the «r-th» section, g_r, depends on q_r (p. 12). In the case of a selector code, where any primary code symbol is expressed always by an equal number of secondary code symbols γ_r for its rank, the following condition must be fulfilled:

$$C_{g_r}^{\gamma_r} \geqslant q_r$$

In case of a binary code, where the primary code symbols may be expressed by different numbers of secondary code sym bols, the corresponding condition is:

$$2^{g_r} \geqslant q_r$$

When a non-local code is used all the fields of the same zone consist of an equal number of bits and are divided identically into sections. The fields of a local code zone may consist of different numbers of bits and be divided differently into sections, small fields not being necessarily divided into sections at all. With the superpositional code there is only one field and a single division into sections.

3. Example of a Punch Card Model

In order to illustrate the various systems of coding we shall consider one imaginary example. Let us assume that the information elements are the medical records of the patients of some large medical institution, say, a tuberculosis or onkological institute. Then zone I (fig. 6) could be used to record, by means of a local code, all «bibliographical» characteristics: year, unit history number, the patient's name, age, sex, nationality, etc., as well as the symptoms and average anatomo-physiological examination data over the entire period of the disorder, which have many mutually exclusive values, such as blood pressure, pulse, blood and gastric juice analysis data, etc. Symptoms of the disorder which take only two values — presence or absence — such as general lassitude, high fatiguability, absence of appetite, disposition to perspire, etc. are recorded using a direct code, in zone II. Zone III is set aside for the names of all the physicians attending the patient, entered by a superpositional code. Zone IV may be employed to record, by means of a non-local code, all the diagnoses (preliminary, clinical and pathologo-anatomical) and the prescriptions and treatment assigned (medicines, operations, etc.).

Any set of rules will do for entering characteristics in the fields, as long as the code is of the selector type. Thus, for instance, fields holding names could be divided into 7 sections each (according to the number of columns) and the corresponding letter of the name could be entered as two holes out of the twelve possible. The number of possible combinations would then equal $C^2_{12}=66$, which is quite enough for recording the 32 letters of the alphabet * (which are identical primary code symbols for

* This refers to the Russian alphabet, which has 32 letters — Tr.

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

all the ranks) and one primary code symbol designating the absence of a letter ($q_r = 33$; $g_r = 12$; $\gamma_r = 2$), as is shown, for instance, in fig. 7. The choice of separate columns of the card as sections is not the most economical method, but has the following advantages:

a) the primary code symbol corresponding to each column can be printed at the top of the column;

b) punching is facilitated, since ordinary card punches (alphabetic) may be used;

c) the punched cards can be arranged in lexicographical order (alphabetic, for instance) with respect to the symbols in one or several columns, by using a conventional sorter.

It would be more economical, for instance (fig. 8), to divide the fields mentioned above into sections of 7 bits each ($g_r = 7$; $\gamma_r = 3$; $C^3_7 = 35 > q_r = 33$). In this case the name indices could consist of the first 12 letters of each name instead of 7, and the sections could be arranged horizontally, each line serving as one section.

Each bit of zone II corresponds to a definite symptom. If a symptom is recorded in the medical record, a hole is punched in the respective bit of the respective punch card. Thus, some cards may have almost all the bits of this zone punched.

Since the cards may have to be sorted according to year and unit history number, the field in which the year is to be recorded, and the section of the field of zone I used for the unit history number may be separate columns. By punching in each column only one position corresponding to the digit in the rank it represents ($q_r = 10$; $g_r = 12$; $\gamma_r = 1$), we can enter any five-digit number, representing the unit history number, and the last digit of any year of the decade under consideration. Obviously, this is also not the most economical way of utilizing the fields. For instance, the field for the patient's age consists of two sections of 5 bits each, which makes it possible to record any number from 0 to 99 by punching two holes in each section $g_r = 5$; $\gamma_r = 2$; $C^2_5 = 10 = q_r$), as is shown in the two variants in fig. 9. The fields for recording the patient's sex and nationality need not be subdivided into sections. The first field consists of two bits, which is sufficient for recording by selector code one of the two possible indices — male or female ($g = 2$; $\gamma = 1$; $C^1_2 = 2 = q$). The second field consists of 12 bits, making it possible to record, depending on the value of γ accepted ($\gamma = 1, 2, 3, 4, 5$ or 6) any number of C^{γ}_{12} indices (the greatest number being $C^6_{12} = 924$). The remainder of zone I, to be used for recording symptoms, is divided up into fields and sections in a similar manner.

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

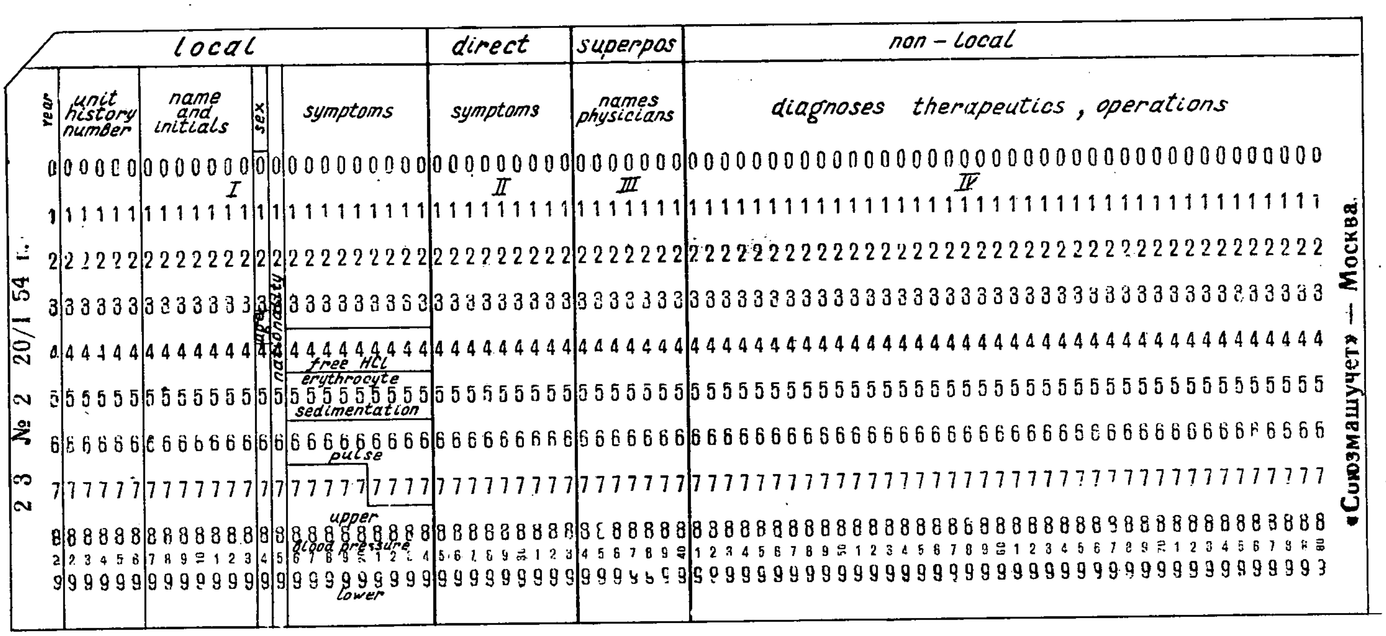
local
direct
superpos
non - local
year
unit history number
name and initials
sex
age
nationality
symptoms
symptoms
names physicians
diagnoses therapeutics , operations
I
II
III
IV
free HCl
erythrocyte sedimentation
pulse
upper
blood pressure
lower
2 3 № 2 20/1 54 г.
«Союзмашучет» — Москва.

Fig. 6

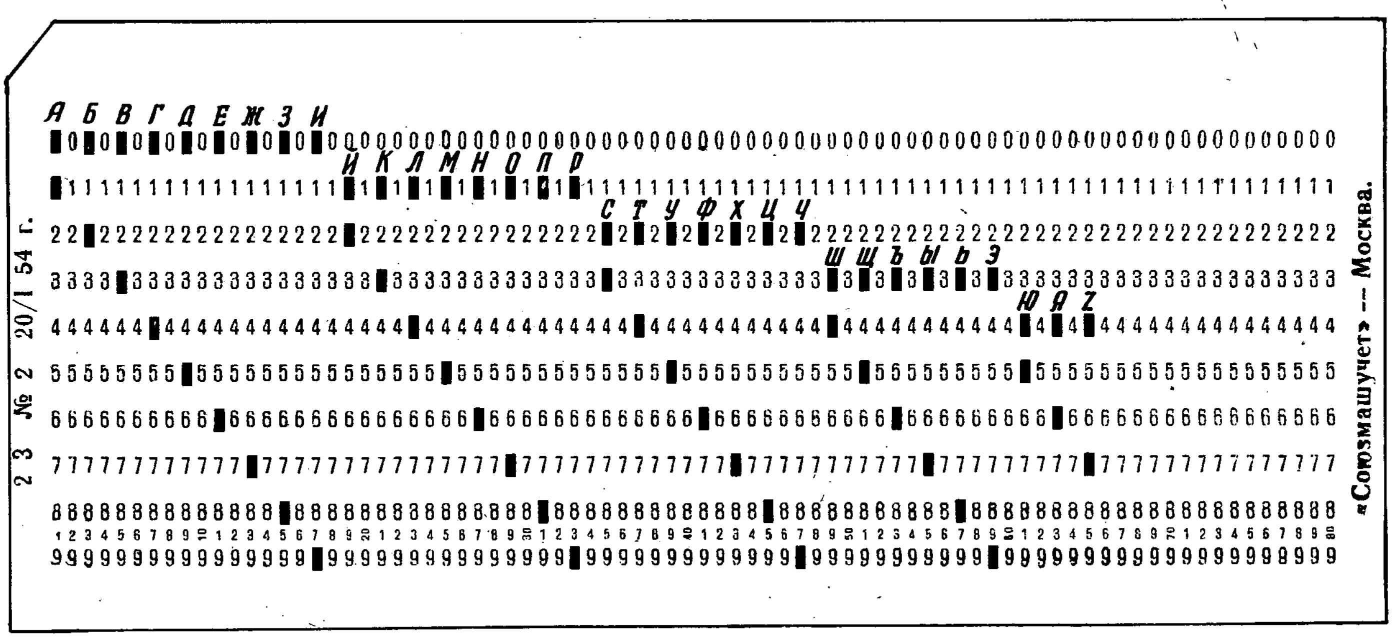
А Б В Г Д Е Ж З И
И К Л М Н О П Р
С Т У Ф Х Ц Ч
Ш Щ Ъ Ы Ь Э
Ю Я Z
2 3 № 2 20/1 54 г.
«Союзмашучет» — Москва.

Fig. 7

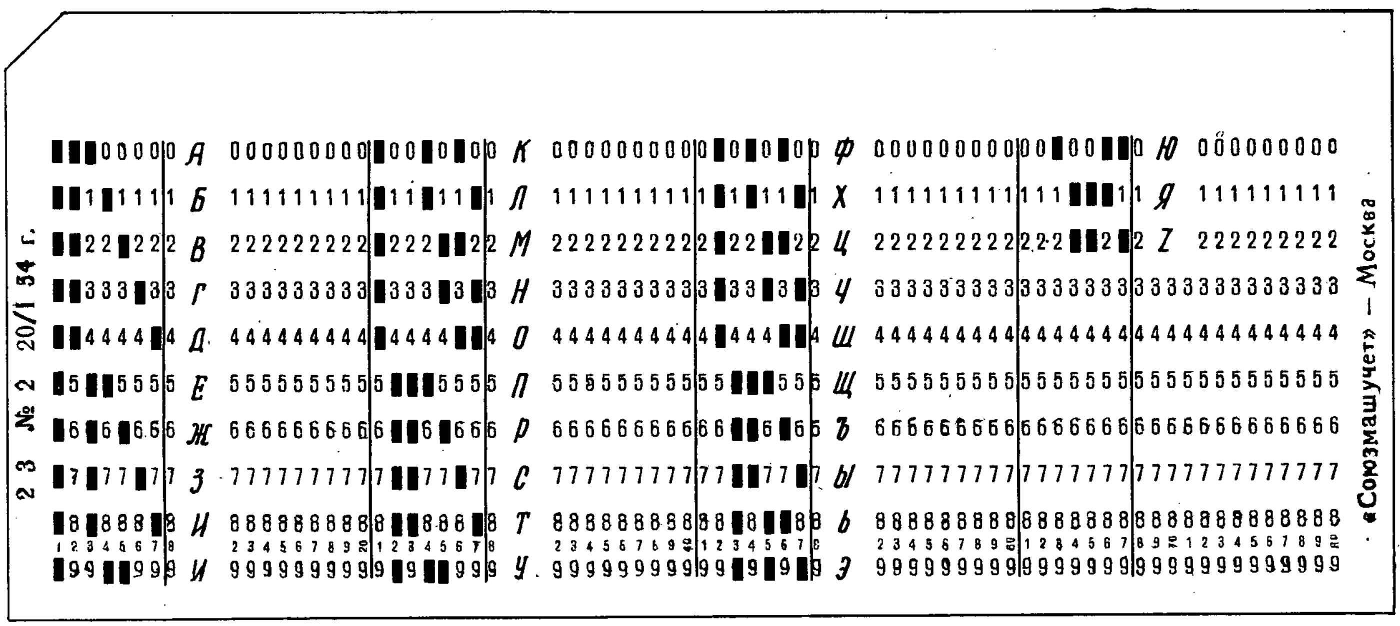

Fig. 8

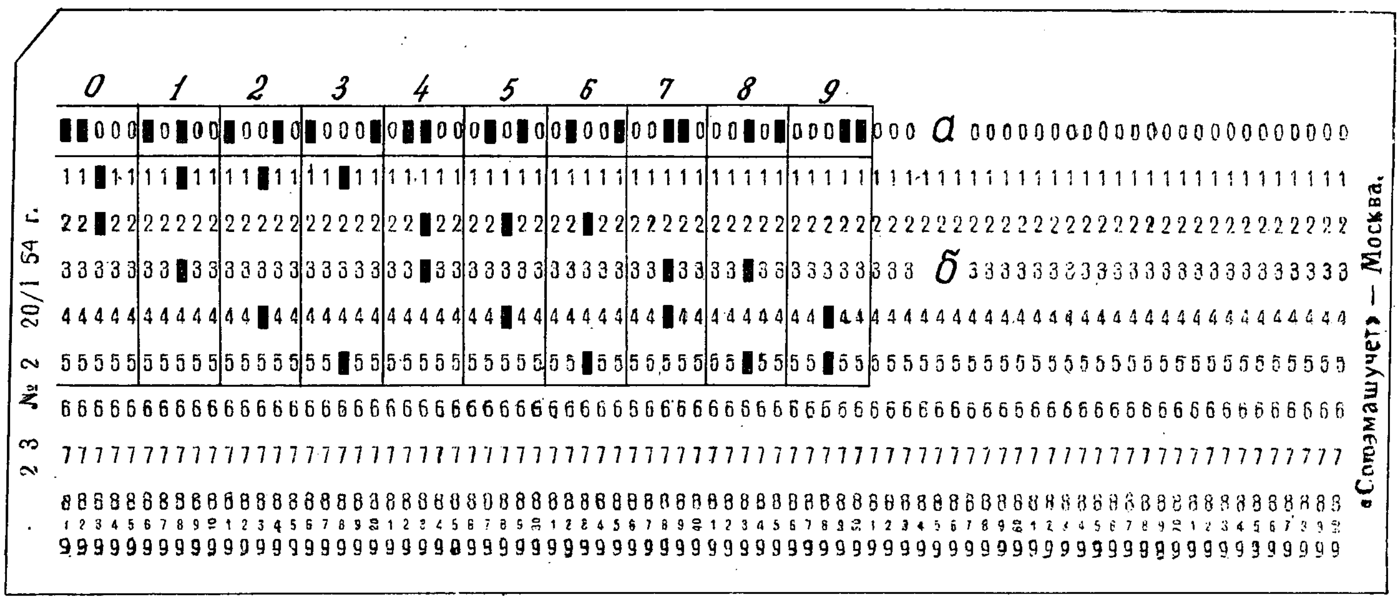

Fig. 9

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

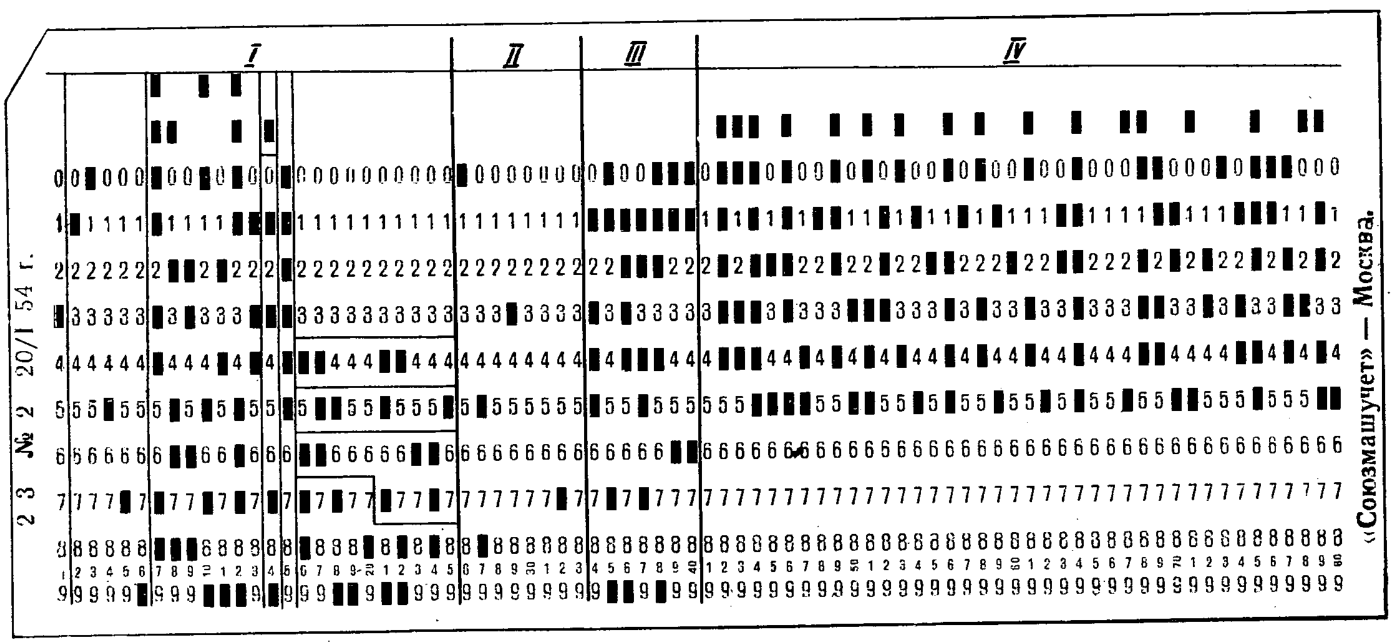

Fig. 10

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

Zone IV consists of 12 fields, arranged horizontally, line by line. This arrangement complies with the above mentioned necessity of scanning the fields consecutively when matching. The fields are broken up into sections (identically this time for all the fields) depending on the non-local code characteristic indices (for diagnoses, operations, medicines, etc.). If the indices in the vocabulary of these characteristics are eight-digit numbers, each field may be subdivided into 8 sections of 5 bits each ($g_r = 5$; $\gamma_r = 2$; $C^2_5 = 10 = q_r$). If they are made up of leters, then, adding two columns to the zone, each field is subdivided into 6 sections of 7 bits each ($g_r = 7$; $\gamma_r = 3$; $C^3_7 = 35 > q_r = 33$). The subdivisions may be more complicated: for instance, the first section may consist of 8 bits, making it possible to record any of 70 primary code symbols which may appear in the first rank of the index ($g_1 = 8$; $\gamma_1 = 4$; $C^4_8 = 70 = q_1$); the second section may have 7 bits for recording any letter of the alphabet, and the remaining 5 sections may be left for recording figures. Finally, a less convenient, but the most economical code may be used, by which, without breaking the fields up into sections, any of C^{20}_{40} indices may be entered by punching 20 holes.

Zone I

Field	Index data	Index	Code
Year	1953	3	C^1_{10}
Unit history number	10579	10579	•
Name and initials	К. Г. Коринфская	К. Г. Коринфская	C^3_7 (fig. 8)
Sex	female	11	C^1_2
Age	58	58	C^2_5 (fig 9b)
Nationality	Russian	0, 1, 2, 3, 4, 5	C^6_{12}
Blood pressure:			
a) upper	135	135	C^2_5 (fig. 9a)
b) lower	70	70	•
Pulse	72	072	•
Erythrocyte sedimentation test	43	43	•
Free HCl in gastric juice	0	00	•
etc.			

Zone II

Index data	Index
Pain	26; 0
Loss of weight	27; 5
Absence of appetite	27; 8
General lassitude	29; 3
Filling deficiency in stomach when X-rayed	32: 7
etc.	

Zone III

Index data	Index	Code
Костина	Костина	C^2_{12} (fig. 7)
Матуленко	Матулен	
Широченко	Широчен	
etc.		

Zone IV

Index data	Index	Code
Achyllic gastritis	LЖ 12437	C^4_8; C^3_7 (fig. 8);
Cancer of the stomach	LЖ 12780	C^2_5 (fig. 9a)
Secondary anemia	QK 57992	
Fibrous pulmonary tuberculosis	RЛ 37865	
Metastasis in the liver	LП 12754	
Sub-total resection of the stomach	L'Ж 12795	
Symptomatic therapy (morphine)	WC 25639	
etc.		

Fig. 10 shows the punch card for a certain concrete (again imaginary) medical record. In order to make clearer the use of the codes described above, let us see how some of the index data of this medical record are encoded.

The indices of the characteristics, together with the headings of the fields (in the case of zone I) and the selected codes, as has been mentioned above, fully reflect the positions of the holes punched on the card. This can readily be seen if the indices in the above tables are compared with the holes punched in the card shown in fig. 10. There are, however, a few details which should fe explained.

Note, first, that in the model the section sequences of the fields of the card, corresponding to the first, second, third, etc. ranks of the index may be arranged either from top to bottom (К. Г. Коринфская, age 58), or from left to right (No. 10579, Er. sed. test, 43, Костина, LЖ 12437, etc.), or in both ways alternately (pulse 072, upper blood pressure 135). Obviously, any type of corrensponedence between ranks and sections may be accepted, as long as it is convenient for the operator and for decoding (reading).

The indices 11 (sex) and 0, 1, 2, 3, 4, 5 (nationality) of zone I denote the position numbers in the respective fields in which the holes are to be punched. Similarly, for zone II the first number in the index denotes the column number, and the second — the line number.

In entering characteristics according to the superpositional code in zone III we come across the above mentioned superpositioning of symbols. Thus, for instance, the holes in this zone make up the name «Локхина» and a number of others. However, this is no predicament in our case, since the probability of physicians with such names working at the institution under consideration is very low. Evidently, the greater the number of characteristics recorded in zone III, the greater will be the number of such superfluous, unnecessary names. Thus, the usually small number of physicians attending each patient is also a favourable factor. These questions can be studied in greater detail by applying the probability theory.

The indices of zone IV are recorded, as is known, in separate lines of the punch card in arbitrary order (on the card under review they are recorded in the same order as the characteristic information in the text). It is assumed that the characteristic information to be entered in this zone has been systematized and is listed together with the corresponding symbols in a special vocabulary. In our example the first rank of the index may be any one of 70 different arbitrary symbols ($C^4_8 = 70$), the second may be any letter of the alphabet, and the rest of the ranks are numerical digits.

Searches of the medical records may be made on any questions or sets of characteristics, both for statistical (e. g. to find the number of Ukrainian women aged between 40 and 49 treated at the institute in 1951 for cancer of the esophagus), and for scientific purposes, for instance, to establish the relationships between symptom complexes and disorders, or to find similar medical records, etc. It would be useless in this article to elaborate on the expediency of the various questions that could be put, since the authors are not specialists in medicine, and have taken the above example, as has been mentioned before, not from practical experience, but just as an illustration *. A similar remark should be made concerning the model described (fig. 6), in which a great variety of field arrangements and hole configurations (codes) have been employed purposely for the sake of illustration. The work of the punch operator would be greatly facilitated and accelerated if a smaller number of codes were used and the symptom fields in zone I arranged more conveniently (for instance, vertically instead of horizontally).

Thus, the final (work) model of the punch card may differ very greatly from the above. For other purposes the models may be entirely different — some of the zones may be missing, or several non-local code zones may be included, the subdivision of the zones into fields and the fields into sections may be different, etc. The main point we should like to emphasize, however, is that the EIM can work with any model at all, without entailing changes in design, since the operation of matching the index data on the punched card with the characteristics in the storage unit is the same, irrespective of the model.

---

* An analogical application of punched cards (using a much simpler model) at an American hospital is described in an article by F. Bradley, C. O. Vermillion, W. Anderson, «Medical Records on Punch Cards». The Modern Hospital, vol. 82, Nos. 4, 5; 1954.

4. Principles of Matching Index Data

Since the location of the fields of different characteristics recorded according to the superpositional, direct and local codes is known, these characteristics can be matched by merely establishing a single valued correspondence between the bits of the punched card and those of the storage unit, and matching the symbol configurations on the punch card with those in the storage unit. In this case all the fields can be matched simultaneously, the results for all the characteristics being registered in a single registering element. When a non-local code is used, the line of the punch card in which the field of any given characteristic may be located, is not known. Therefore, in this case each of the characteristics on the card must be matched consecutively with all the characteristics in the storage unit. The results of individual matchings made in the course of one run of the card are registered separately for each characteristic of the storage unit by means of its own registering element. Since each pair of matched fields have equal dimensions, a single valued correspondence is established between the bits of these fields also, and it is again the symbol configurations that are matched.

As has been mentioned above, with superpositional and direct codes the symbol configurations are matched for inclusion, i. e. it is established whether the plurality of holes in the card includes the given hole plurality corresponding to the plurality of code symbols in the storage unit. In the case of local and non-local codes they are matched for coincidence of pluralities on the fields matched. Matching for inclusion is much simpler with all devices than matching for coincidence. It can easily be seen, however, that when using a selector code the results of matching for inclusion are identical with those of matching for coincidence, since the pluralities matched always contain the same number of holes. This is why only selector codes are used in the EIM machine.

Now let us see how characteristics of superpositional, direct and local code zones are matched. We shall assume that these three zones take up a part of the punched card consisting of a certain number of complete columns. Taking it for granted that the fields of the storage unit in which question data have not been specified (or even the sections in which primary code symbols have not been specified) do not contain secondary code symbols, we shall define the selection specification for the cards required more accurately as follows: those cards are to be selected whose plurality of punched bits M (in the part of the card under consideration) includes the plurality of bits containing code symbols in the storage unit M* (the bits of the punch card and the storage unit being identified).

Making use of the example discussed in the previous paragraph we shall explain these selection specifications for two cases when they are and when they are not fulfilled. In the first case let the search be made on the following question: «to find all the medical records for Russian females aged from 50 to 59 with no free HCl in their gastric juice, with erythrocyte sedimentation test showings from 40 to 49, with absence of appetite and with filling deficiency in stomach when X-rayed, who suffered in 1953 from achyllic gastritis and secondary anaemia and were subjected to subtotal resection of the stomach by physician Широченко». The question data are entered in encoded form in the storage unit in accordance with the accepted card model (fig. 6). A punched card containing holes corresponding to the code symbols in the storage unit, is shown in fig. 11. In the second case let the inquiry differ from the first only in dealing with men subjected to total resection of the stomach (index — L'Ж 1296). Hence, the punched card for the second question, which is shown in fig. 12, differs from the first only in that the 12th position of column 14 is punched instead of the 11th, and that the hole of the lower characteristics in the last section is in column 80 instead of 79.

If we now check the fulfilment of the selection specification for the card shown in fig. 10, ignoring the right half of the card, then by matching the plurality of holes (M) contained in the first 40 columns of the card with the plurality M * contained in the first 40 columns of the cards shown in figs. 11 and 12, it will be seen that in the first case M includes M *, whereas in the second case it does not. Hence, for the data specified by the first question and recorded in the first three zones, the punched card is selected, while for the data of the second it is not, in accordance with the meaning content of the two questions.

Since the scanning unit of the EIM machine consists of 80 bruches, under which all the 12 lines (positions) of the punched card pass consecutively, the fulfilment of the selection specifications is checked in the course of 12 strokes. During the «i-th» stroke the bit pluralities M_i and M_i^* of the «i-th» line of the punched card and the storage unit will be matched. The matching of the two lines is based on the easily proved thesis that the inclusion of M_i^* in M_i is equivalent to non-intersection between the pluralities $\bar{M}_i$ and M_i^*, where $\bar{M}_i$ is the plurality of unpunched bits in the line, supplementary to the plurality M_i. In order to find whether these pluralities intersect or not, all the brushes in the storage unit corresponding to the bits of M_i^* should be connected to the registering element, and the wiring should be set up so that at the moment of matching current pulses could go through to the registering element only from brushes which do not make contact with the contact roll through a hole punched in the card, i. e. from brushes corresponding to the bits of $\bar{M}_i$. Then, if M_i^* and $\bar{M}_i$ intersect, the registering element will be energized, while if they do not intersect, no current pulse will be sent to the registering element, and the latter will not be energized. If the registering element is not energized in any of the 12 strokes, the pluralities $\bar{M}$ and M^* have not intersected, i. e. M includes M^*, and the selection specifications are fulfilled.

Thus, for instance, for the first question (fig. 11) all the M_i shown in fig. 10 include M_i^*, and the registering element will not be energized at all. For the second question only M_{12} does not include M_{12}^* ($\bar{M}_{12}$ intersects with M_{12}^* only in one bit in column 14), but this is enough for the registering element to be energized and register non-fulfilment of the selection specifications.

The main difficulty in carrying out such a matching operation is the necessity of evolving, i. e. matching one by one, the lines of the storage unit (the lines of the punched card are evolved automatically as the card moves under the brushes). On the basis of the foregoing, the introduction of any line of the storage unit into the matching operation reduces to connecting the brushes corresponding to the bit plurality M_i^* in the «i-th» stroke to the registering element. The use of a large number of relays, selectors, etc. for this purpose involves a very large number of contacts (up to 960), while the use of a question card rotating past a second set of brushes would require substantial changes in the design of the sorter on which the EIM machine is based. Besides, such a solution would be of no use, as we shall see further on, in matching index data recorded by a non-local code. The method accepted for the EIM machine consists in joint interplugging of the brushes into groups corresponding to each plurality M_i^*, each group being subsequently switched to the registering element by means of only one cam contact or relay. Obviously, the brushes must be interplugged through intermediate elements, since connecting the brushes directly would give rise to by-pass circuits.

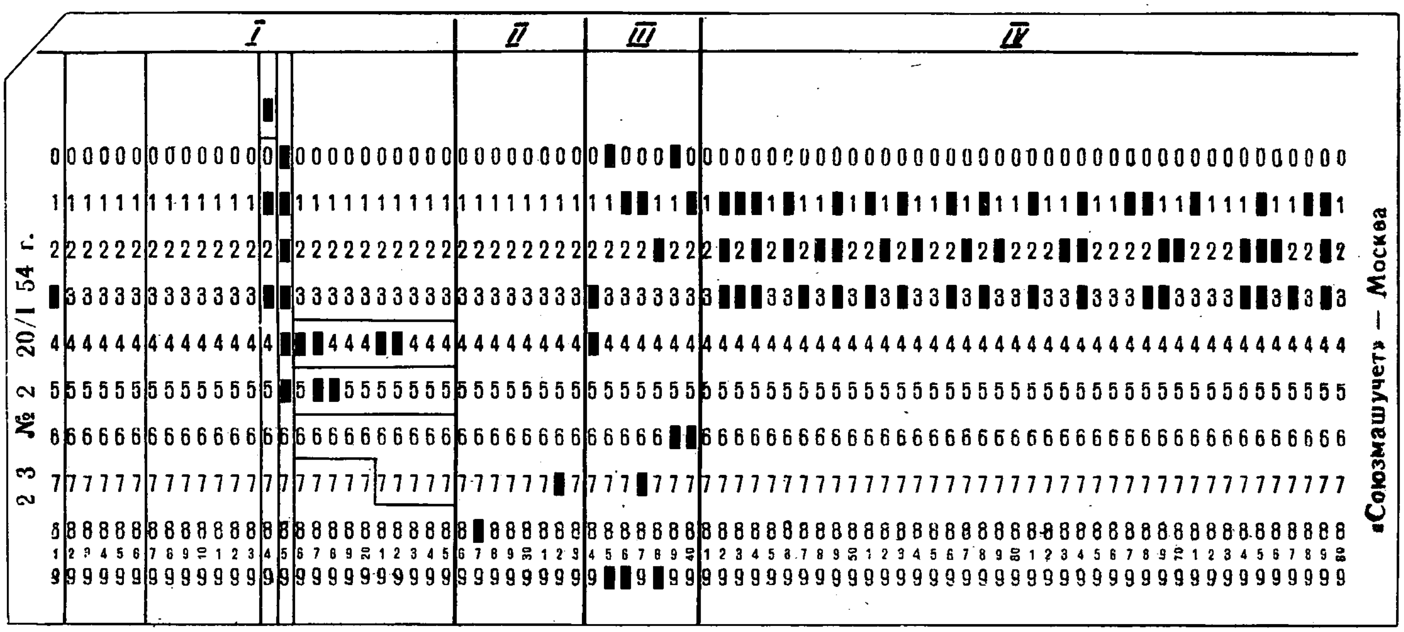

Fig. 11

In case of a non-local code zone consisting, as before, of a certain number of complete columns, each line of the punched card (rather, the part of the line belonging to the zone in question) must be matched not only with the corresponding line of the storage unit, but with all its lines. In matching each pair of lines—the «i-th» line of the card ($i=0, 1, ..., 9, 11, 12$) and the «j-th» line of the storage unit ($j = 1, 2, ..., K$, where K is the number of characteristics in the question) — the inclusion of the plurality Mj^* of bits containing code symbols in the «j-th» line of the storage unit, in the plurality M_i of punched bits in the «i-th» line of the punch card is checked, as before. The matching is carried out in exactly the same way as described above, by determining whether the pluralities M_i and Mj^* intersect. Thus, if the registering element is not energized when M_i and Mj^* are matched, M_i includes Mj^*, i. e. the characteristic recorded in the «j-th» field of the storage unit coincides with the «i-th» characteristic of the punched card.

Since the specifications for the selection of the desired cards in the case of a non-local code are of a more complex nature than the simple requirement of inclusion of all the characteristics specified in the storage unit within the aggregate of index data entered on the card, each characteristic in the storage unit has its own assigned element which registers the inclusion of that characteristic in the index data on the card. This also solves the probem of evolution — the matching of all the lines of the storage unit with each line of the card. Each line of the card is matched in this case at once with all the lines of the storage unit simultaneouly interplugged through intermediate elements with the groups of brushes corresponding to the pluralities Mj^* and with their individual registering elements. During each run it is thus found which of the registering elements failed to be energized for at least one line of the punched card, thus revealing which of the characteristics in the storage unit are contained among the index data on the punched card. If the number or relative position of the registering elements not energized complies with the selection specifications, which is determined by means of the selecting unit, adjusted for these specifications, the card will be selected.

This will become clearer if explained for our example. The index data of the non-local code zone shown in fig. 10 is represented by the punched bit pluralities M_0, M_1, ..., M_5, M_{11} (designated according to their respective line numbers). The question data (fig. 11, 12) is represented by the pluralities M_1^*, M_2^* and M_3^* (numbered by the line from top to bottom). For the first question (fig. 11) among the pluralities M_i shown in fig. 10 the pluralities M_{11}, M_1, and M_4 include the pluralities M_1^*, M_2^*

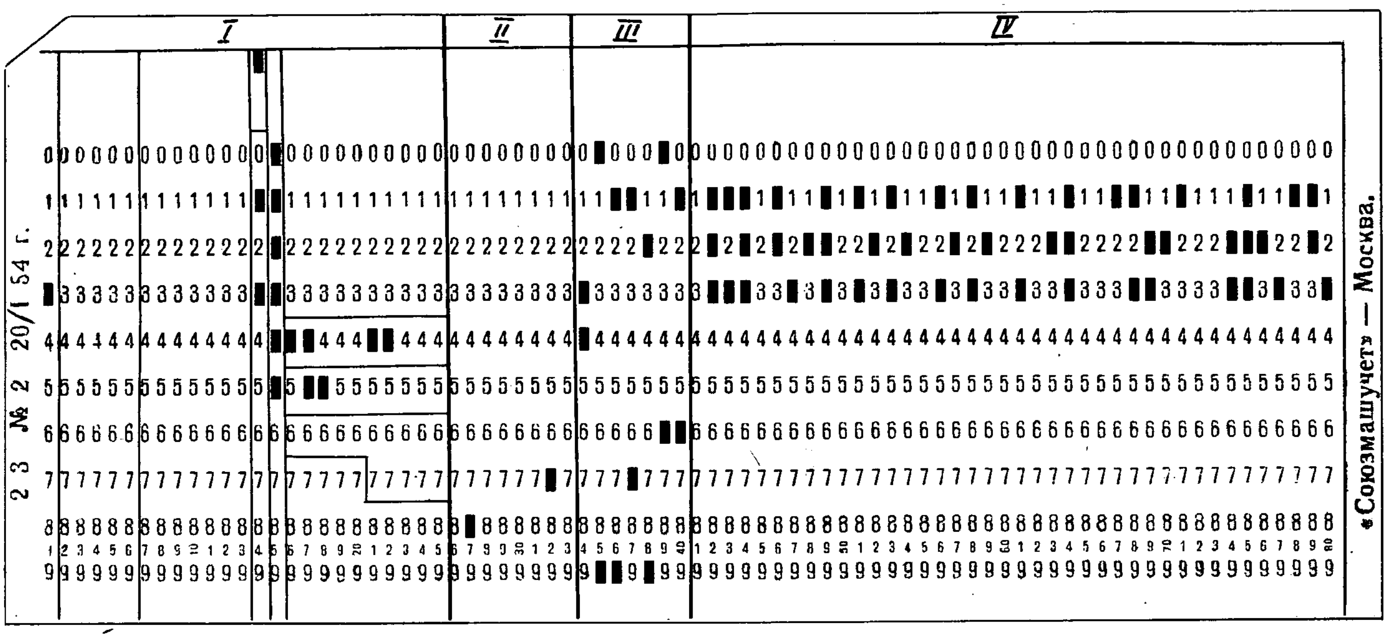

Fig. 12

and M_3^* respectively. Therefore, when line 11 of the card comes under the brushes the registering element of the first question characteristic will not be energized; when line 1 comes through, the element of the second question characteristic will not be energized, and when line 4 goes through, the element of the third characteristic will not be energized. Since the selection specifications in our example consist merely in the requirement that all three question characteristics be included in the index data on the punched card, these specifications are fulfilled in this case and the card will be selected. For the second question (fig. 12) the plurality M_3^* is not included in any of the pluralities M_i', and therefore the registering element of the third characteristic will be energized for all the lines of the card, which will register non-fulfilment of the selection specifications.

There will obviously be no difficulty in adjusting the wiring of the selecting unit to simultaneously take account of the condition of the registering elements for several non-local code zones and for superpositional, direct and local code zones, making it possible to select the punched card in one run according to all the selection specifications. If the capacity of the storage unit is increased, it can be made also to hold several questions at once and to select in one run all the cards complying with at least one of the questions. This greatly increases searching speeds and means less wear and tear for the cards. The selected cards are then examined separately on each question.

---

5. Recording and Reading Information

The coded designations of the index data of the information element are entered on a punch card by an operator using a card punch.

When the index data is encoded by superpositional, direct or local codes, to be entered as holes punched according to a definite pattern on the punch card, a conventional 80-column card punch is used. The design of this punch is slightly modified to permit punching various numbers of holes in the same column of the card (fig. 13).

When the index data are encoded after a non-local code, according to which each characteristic is entered in one of the positions of the card, two cases of recording are possible. In the first, when the number of different primary code symbols is large, recording is done by means of an ordinary card punch. In the second case, when the number of different primary code symbols does not exceed 40 or 50, the card punch is equipped with a simple device for encoding the individual symbols after the accepted code. When recording, the card passes under the punches of the punch mechanism several times, according to the number of characteristics to be entered.

Another method may also be used for entering index data by non-local code, based on the utilization of the duplicating mechanism of the punch. In this method a master card is prepared beforehand for each characteristic, the latter being punched in a definite position on the card as a combination of holes. The master cards corresponding to the index data of the information element are placed one by one in the duplicating mechanism of the punch and the holes in them are duplicated in various lines of one punch card.

The data punched on the cards are verified by another operator making a second set of cards, after which the two copies are compared. When using the duplicating mechanism the cards are verified automatically.

Approved For Release 2008/04/25 : CIA-RDP80T00246A004400020002-9

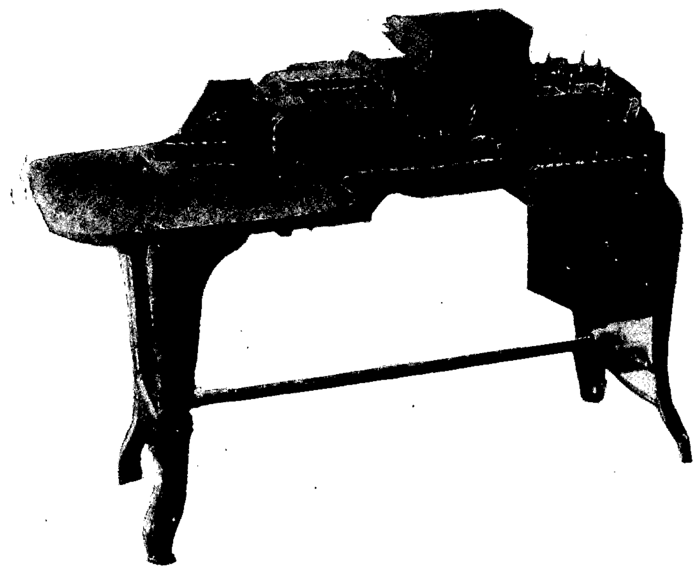

Fig. 13

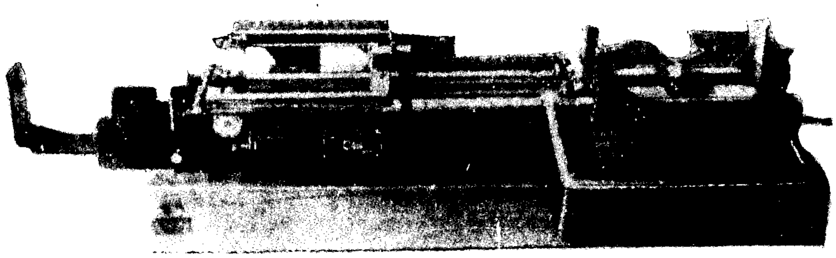

Fig. 14

A conventional verifier (fig. 14) may be used to verify the data entered on the card, if the wiring of the verifier is altered to allow each column to be verified in more than one stroke. Only one hole in the column is verified at each stroke.

After selecting the punched cards in accordance with any specified question, the basic bibliographical on the data information elements contained in them have to be read. These data are, for instance, the catalogue number or some other conventional number. These data being entered on the card in the form of punched holes, they have to be listed on paper to be read.

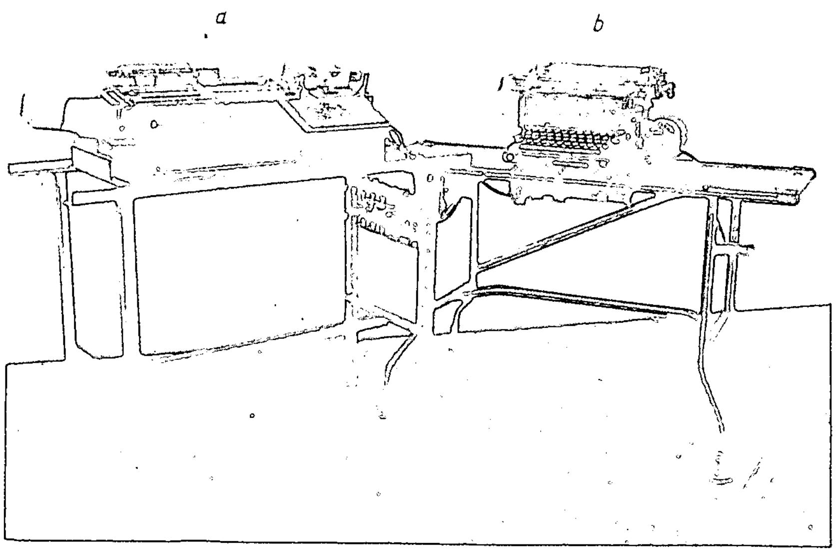

Fig. 15

Listing from the punched cards may be done by means of an ordinary or alphabetic tabulator. However, inasmuch as the bibliographical data are compact and the time needed to list them, say, with a typewriter, is not very great, a comparatively inexpensive listing device is used for this purpose. This device for listing data from punch cards consists of a motor-driven typewriter and a conventional punch card verifier (fig. 15). The two units are electrically connected.

The punched cards, the data of which have to be listed, are placed in the magazine of the feed mechanism of the verifier (fig. 15-a). When listing, the card passes under the scanning mechanism of the verifier column by column, and the data are transmitted as electrical pulses to the magnets of the typewriter (fig. 15-b), which lists the data automatically on paper.

6. EIM Design

The EIM machine (fig. 16) was built on the basis of a C 80-1 horizontal sorter and is an improved modification of the latter. The improvements, intended to impart to the EIM, the

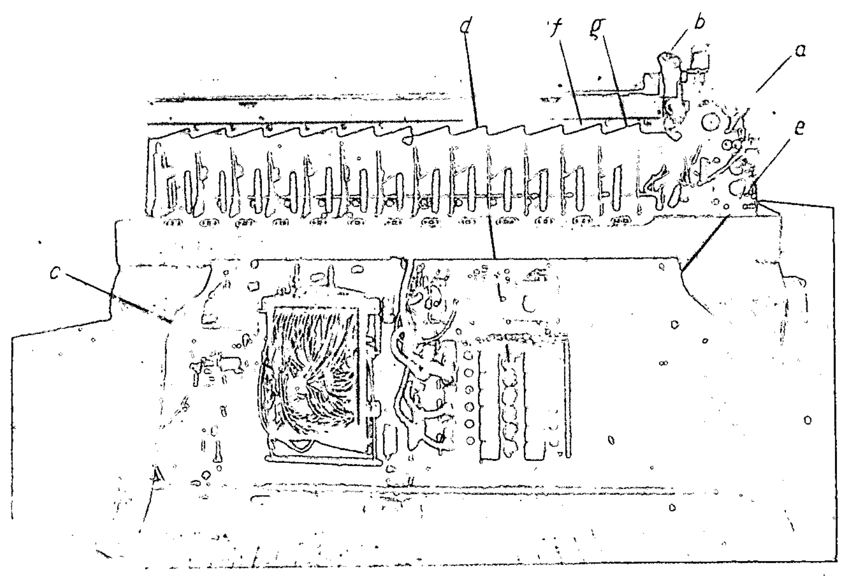

Fig. 16

required universality in fulfilling its taks, involved alterations in the scanning mechanism and in the mechanism for entering question data (storage unit); besides, an electrical wiring diagram was developed and set up for automatic selection of the required information.

The wiring diagram for automatic selection of information required consists, essentially, of two sections. The first section is the diagram for matching the information element characteristics. The second is the diagram for registering the results of the matching operations and controlling the device for transporting the punched cards to the respective pocket of the machine.

The matching diagram has no electron tubes or relays and is made up of inexpensive and reliable elements. The registering diagram includes several electronic and relay elements, the number of which depends on the number of characteristics in the question. The number of elements in the part of the registering diagram which serves the matching diagram for characteristics recorded by a non-local code must be equal to the number of characteristics in the question. The part of the registering diagram which serves the matching diagram for characteristics entered on the punch card in other codes requires only one element. Such electrical wiring for automatic selection of information provides the required universality of the EIM operation at low cost and with sufficient reliability.

As has been mentioned above, the EIM can select information entered on punch cards in codes of any kind, or combinations of codes, except the ordinal code. Especially noteworthy is the possibility of selecting information simultaneously for several question characteristics in one run of the batch of cards to be examined. The convenience of this method of selection is in the possibility of introducing various specifications of logical relations between the characteristics of the question. Special electronic units were developed and included in the wiring diagram of the EIM, permitting the introduction of logical specifications between any 4 question characteristics entered in a non-local code.

The additional units include an electronic counter for registering the number of question characteristics (from 1 to 4) which coincide with the characteristics of the information element examined. The capacity of the counter can be increased if necessary. By using this counter information elements containing a certain predetermined number of question characteristics can be selected.

An electronic selecting device has also been developed, by means of which information elements containing, for instance, the characteristics A and B or C, but not containing D, etc. (page 10) can be selected.

It should be noted that selection of information entered in a non-local selector code, if the question contains only one characteristic, can be done, using a very simple matching diagram.

The principles of design of the EIM wiring diagram make it possible, if necessary, to increase the capacity of the storage unit 5 to 8 times without great difficulty, thus permitting information to be selected simultaneously for several questions.

Like the conventional horizontal sorter, the EIM has a mechanism for feeding cards at a rate of 400 per minute, and a magazine, in which the cards to be examined are placed (fig. 16-a).

When the EIM is set into operation the cards are fed one by one from the magazine to the brushes of the scanning unit (fig. 16-b). Unlike the ordinary sorter, which has only 1 or 12 (machine with a selector device) brushes, the scanning unit of the EIM has 80 brushes, one for each column on the punch card.

The question data are set up on a removable plugboard (fig. 16-c). The same plugboard is used to set up the specifications for selecting the required cards. A more convenient storage unit of greater capacity has been designed in the form of a keyboard.

Fig. 16-d shows the registering elements of the wiring diagram, which register the results of matching the cards. There are 12 elements for non-local codes, and one for other codes.

The greater part of the EIM wiring is mounted on 2 special panels (fig. 16-e).

The selected cards pass into pocket I (fig. 16-f), and the rest go to pocket II (fig. 16-g).

Translated by D. Sobolev.